Small Objects Detection Based on Multi-scale Feature Fusion

Guozhen Wang*
Department of Computer Science and Technology, Tianjin University Renai College, Tianjin, China
*Corresponding author(E-mail:hangdaichu@126.com)

Abstract
Small objects detection is a hot research direction in the field of object detection. We propose an improved method to solve the problem that the classical object detection algorithm SSD has poor detection performance on small objects. We use oversampling strategy to balance the sample numbers of small and large objects, allowing more small objects to enter training. Using multi-scale feature fusion, fuse semantic information contained in low-resolution feature maps into high-resolution feature maps, we can get more location and category information of small objects when predicting. Experiments show that multi-scale feature fusion using oversampling strategy can improve mAP of small objects detection. We obtain a mAP of 75.2% on the PASCAL VOC2007 dataset, and achieve 0.9% improvement compare to SSD. Our model not only has a great improvement on small objects detection, but also has different degrees of improvement on the detection of medium and large objects.

Keywords: Object Detection, Multi-scale Feature Fusion, Oversampling Strategy

1. Introduction
Object detection has always been the hottest research direction in the field of computer vision. Traditional object detection uses manual design to extract feature factors and then classify them using classifiers to achieve object detection. With the maturity of convolutional neural network technology and the widespread use of hardware devices such as GPUs, object detection based on convolutional neural networks has achieved excellent performance in object detection and achieved higher and higher average accuracy. Object detection based on convolutional neural networks has the incomparable advantages of traditional object detection, and more and more researchers turn to deep learning.

At present, there are two main object detection algorithms. The first is based on region proposal networks. Such algorithms are represented by R-CNN, Fast R-CNN, and Faster R-CNN [1]. Proposed regions where target objects may exist are selected, and then feature extraction and object classification are performed on these proposed regions. Such algorithms have problems of large computational complexity and poor real-time performance. The second is regression-based object detection algorithm, represented by YOLO [2] and SSD[3]. This type of algorithm can directly return the position information and target category information at multiple positions on the image, and the detection speed is fast, but the accuracy is not as good as the object detection algorithm based on proposed region. Both of these algorithms have good accuracy in medium and large objects detection, but they are not satisfactory in the accuracy on small objects detection. In fact, small objects detection is even more important in practical applications. In self-driving, identifying and detecting distant vehicles is important for safe driving. In high-resolution images, pedestrians and traffic signs account for only a small percentage of the entire image. In satellite remote sensing images[4], buildings, trees, cars and other objects have few pixels and are almost invisible. Therefore, small objects detection has more important research value in the real world.

Based on the SSD, this paper uses the fusion of multi-scale features to enhance the detection of small objects, and to improve the accuracy of small objects detection. First, we have oversampled the image containing the small objects for data augmentation [5]. Then we use deconvolution layers to fuse feature maps of different scales. The fused feature map is used for regression, and the small objects are detected, and finally the position information and category information of the small objects are obtained. Experiments show that on the PASCAL VOC dataset, not only the accuracy of small objects detection is improved, but also the detection accuracy of medium objects and large objects are improved.

2 Related Work
A lot of researchers have done lots of works to improve the accuracy of small objects detection. Wen [6] et al. use Atrous convolution to improve the resolution of the fused feature map. The original image is sampled in 7 ratios for data augmentation, the sensitivity of the small objects area is enhanced, and the accuracy of small objects detection is improved. On the basis of SSD, Cao [7] et al. introduced the semantic information of high-level feature maps into low-level feature maps by using multi-layer feature fusion to improve the accuracy of small objects detection. Xing [8] improved the SSD network architecture and adjusted the scaling factors of

76
different feature layers according to the object size. Better accuracy is achieved with small pedestrian size. Ren [9] et al. used random rotation for data augmentation, using sampling strategy to balance the number of images for different categories, and then using a single high-resolution feature map combined with context information to further enhance the small objects detection performance. Li [10] proposed Perceptual Generative Adversarial networks. In this architecture, the representation of small objects is promoted to super-resolved objects, narrowing the representation difference between small objects and large objects to improve the detection performance of small objects, and achieving good accuracy in traffic signs and pedestrians.

3 Small Objects Detection Based on Multi-scale Feature Fusion

3.1 Problems of small objects detection

In general, in an image, we call an object smaller than $32 \times 32$ pixels a small object, an object larger than $32 \times 32$ and smaller than $96 \times 96$ pixels is called a medium object, and an object larger than $96 \times 96$ pixels is called a large object. Regardless of the proportion of the area occupied by the objects in the image, or the number of pixels owned by the objects, the small objects is much lower than the large objects. Most object detectors currently prefer large objects and medium objects.

The detection accuracy of SSD in small objects is much lower than the accuracy of large objects. SSD uses multiple feature maps of different scales to detect the object separately, and finally combines all the intermediate results to obtain the final result. Its calculation process can be abstracted into the following mathematical formula [11].

$$\begin{align*}
F_n &= \sigma_n(F_{n-1}) = \sigma_n(\partial_n(\cdots \partial_1(I))) \\
R &= D(\sigma_n(F_n), \cdots, \sigma_{n-s}(F_{n-s})), n > s > 0
\end{align*}$$

Where $F_n$ represents the nth layer feature map, $\sigma_n$ represents the nonlinear transformation from the n-th feature map to the n-th feature map, such as combination of convolutional layer, pooling layer, and activation function. $\partial_1(I)$ represents the first feature map obtained by the inputting image I. $\sigma_n(\cdots)$ represents the intermediate detection result of the corresponding feature map. $D(\cdots)$ represents the operation of merging all intermediate detection results. R represents the final detection result.

From formula (1) and formula (2), we can see that the SSD model is a "pipeline" shape, that is, the output of the previous layer in the network architecture is used as the input for the next layer in the network architecture. The shallow feature map in the SSD has high resolution and contains more detailed information and position information, which can help better locate the object; deep feature maps have low resolution and contain more abstract semantic information. The feature map of each layer is only responsible for the detection of a specific size, and requires sufficient detail information and semantic information to accurately locate and classify the object. After a series of convolution and pooling operations, the small objects leaves less and less information on the high-level feature map, and may even be blank. Therefore, SSD mainly uses low-level detail features to detect small objects, but the low-level feature maps used for small objects detection are only conv4_3, and the feature expression ability is not enough, resulting in poor detection performance on small objects.

On the other hand, the smaller the small objects is in the whole image, the less annotated pixels belong to the small objects. In an image, only a few percent of the annotated pixels belong to the small objects and most of the annotated pixels belong to medium and large objects. SSD generates different default boxes at each detection layer, and selects the default box with an IOU greater than 0.5 as the positive sample, and the default box with the IOU less than 0.5 as the negative sample. It is obviously more advantageous for large objects. In the training process, large objects often get higher IOU, matching more default boxes, and the number of positive and negative samples is balanced. The small objects IOU value is very low, and the default box with IOU greater than 0.5 is less, resulting in an imbalance in the number of positive and negative samples of small objects. This leads to more information about the large objects learned by the detector during the training process, and it is impossible to learn enough small objects information. This is also a major reason for the poor detection performance on small objects.

3.2 Oversampling

On the PASCAL VOC dataset, the distribution of small objects and large objects is imbalance. The number of images containing small objects is much smaller than the number of images containing medium and large objects. That is to say, during the training process, the object detector can obtain enough information of medium objects and large objects and obtain good detection performance. However, there is not enough small objects enter the training, so that the object detector can only obtain little information about small objects, which also directly leads to unsatisfactory performance on the small objects detection. In order to solve the problem of imbalanced samples between small objects and large objects, we use oversampling strategy to oversample images containing small objects. Through the oversampling strategy, the sample numbers of small objects is increased, and the small objects get data augmentation. Using the oversampling strategy for data augmentation
is a simple and straightforward method. The object detector can also learn more information about the small objects during the training process, thereby improving the detection performance on small objects.

3.3 Multi-scale Feature Fusion

SSD only uses conv4_3 to detect small objects, and cannot obtain enough location information and semantic information. We introduce a deconvolution structure, through the deconvolution operation to achieve the mapping of low-resolution feature maps to high-resolution feature maps, so that low-resolution high-semantic feature maps and high-resolution low-semantic information feature maps are fused to improve detection performance on small objects.

The effect of deconvolution is reverse to the convolution operation, mapping low resolution feature maps to high resolution feature maps. Deconvolution works the same as oversampling, but is different from oversampling with fixed parameter values. The process of deconvolution is the inverse of the forward propagation of convolution in convolutional neural networks. The position of input feature map and the output feature map are interchanged. So, the convolution kernel parameter of deconvolution is the same as the convolution kernel parameter of convolution. And the parameters can be learned and adjusted through back propagation during the training process to make the parameter more reasonable.

The deconvolution operation can make the low resolution high semantic information feature map the same size as the previous high resolution low semantic information feature map, and concatenate the two feature maps into a multi-channel feature map. Convolution operation to extract features from multi-channel feature maps to achieve feature fusion. The parameters of the convolution kernel can be learned and adjusted during the training process, so the feature fusion using multi-channel convolution is more effective than directly adding the feature map to the feature map to achieve fusion. The fused feature map has higher resolution and position information of small objects, has rich semantic information to help learn the categories information of the small objects. The multi-scale feature fusion process is shown in Figure 1.

\[
O = S \times (L - 1) + H - 2P
\]

\[ (3) \]

In the formula, \( O \) represents the size of the deconvolution output feature map; \( S \) represents the stride; \( L \) is the input feature map size; \( H \) is the convolution kernel size; \( P \) represents the padding.

3.4 Feature fusion module

In the feature fusion process, we first add a deconvolution layer behind the feature map with low resolution to increase the resolution of the feature map. A convolution layer is followed to extract feature, then followed by a Batch Normalization layer, and an activation function layer. The convolution layer is added to learn better features for fusion. The purpose of adding the Batch Normalization layer is to make the feature vector satisfy the Gaussian distribution without losing information, and it will not be too large or too small in the gradient propagation process. It can prevent the gradient disappearance and the gradient explosion. We choose ReLU as the activation function. The feature map with large resolution is directly followed by a convolution layer, a Batch Normalization layer, and an activation function layer. The two processed feature layers are then concatenated together. Finally, the concatenated feature map is processed with a 1×1 convolution layer to achieve feature fusion while reducing the dimension. The feature fusion module is shown in Figure 2.
4. Result Analysis

The experimental configuration used in this paper is Intel(R) Core (TM) i7-7700HQ, the memory is 8G, and the GPU is NVIDIA(R)GTX(R)1080TI. The dataset is PASCAL VOC2007 and 2012.

4.1 Oversampling

The first experiment was mainly used to verify whether the oversampling strategy helps in object detection. During the experiment, we manually created a copy of the image containing the small objects instead of stochastic oversampling, which improved efficiency while avoiding the over-fitting problem caused by oversampling. We used different oversampling ratios of 2x, 3x, and 4x to observe the effect of different oversampling ratios on detection performance. Baseline is SSD300.

<table>
<thead>
<tr>
<th>Ratios</th>
<th>mAP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD300</td>
<td>74.3</td>
</tr>
<tr>
<td>Oversampling 2x</td>
<td>74.5</td>
</tr>
<tr>
<td>Oversampling 3x</td>
<td>74.6</td>
</tr>
<tr>
<td>Oversampling 4x</td>
<td>74.4</td>
</tr>
</tbody>
</table>

From Table 1, we can see that compared with SSD, no matter which ratios of oversampling is worked, there is a certain improvement in the average accuracy of object detection. This is because the number of objects with different sizes is balanced by oversampling, and more objects enter the training phase, improving the detection performance. We can see that 3x oversampling performance improvement is the biggest. In the following experiments, we all use 3x oversampling.

4.2 Multi-scale Feature Fusion

We train our model on the PASCAL VOC2007 and 2012 datasets and test on the PASCAL VOC2007 dataset. We use stochastic gradient descent. The learning rate is initialized to 0.001, the momentum is set to 0.9, and the batch size is 8. We used a random initialization parameter strategy, so we increased the number of iterations during training. The training process starts with the learning rate at 10^{-3} for the first 60K iterations, and then 10^{-4} for 20K iterations,10^{-5} for 10K iterations. The IOU threshold greater than 0.5 is set to a positive sample, otherwise it is set to negative sample.

In the multi-scale feature fusion experiment, we fuse the three feature maps Conv3_3, Conv4_3, and Conv5_3 with different combinations to find the fusion method that can obtain the best performance. Then implement the oversampling strategy for the best fusion method and compare the performance improvement effects.

<table>
<thead>
<tr>
<th>Feature map</th>
<th>mAP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv4 3</td>
<td>74.3</td>
</tr>
<tr>
<td>Conv3 3+Conv4 3</td>
<td>74.7</td>
</tr>
<tr>
<td>Conv4 3+Conv5 3</td>
<td>74.5</td>
</tr>
<tr>
<td>Conv3 3+Conv4 3+Conv5 3</td>
<td>74.8</td>
</tr>
<tr>
<td>Conv3 3+Conv4 3+Conv5 3+Oversampling</td>
<td>75.2</td>
</tr>
</tbody>
</table>

Conv4_3 in the table is the feature layer used by the SSD algorithm to detect small objects. From the average accuracy in the table, the use of multi-scale feature fusion is helpful in small objects detection, which is higher than the single-layer Conv4_3 in detection accuracy. Different combinations of feature layers have different detection effects. Conv3_3, Conv4_3, Conv5_3 three-layer feature map fusion get the best detection performance, mAP is 74.8%, 0.5% higher than SSD. The Conv3_3 and Conv4_3 two-layer fusion get a better mAP than the Conv4_3 and Conv5_3. We introduced the oversampling strategy into the network model with Conv3_3, Conv4_3, Conv5_3 and obtain the best detection performance, with mAP of 75.2%, which is 0.9%
higher than SSD. We compare our algorithm with the classic object detection algorithm Faster R-CNN and SSD from mAP, as shown in Table 3. In the table, ours represents a model that uses multi-scale feature fusion with Conv3_3, Conv4_3, and Conv5_3, and use the oversampling strategy.

### Table 3. Results on PASCAL VOC2007

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>TV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>73.2</td>
<td>76.5</td>
<td>79.0</td>
<td>70.9</td>
<td>65.5</td>
<td>52.1</td>
<td>83.1</td>
<td>84.7</td>
<td>86.4</td>
<td>52.0</td>
<td>81.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSD</td>
<td>74.3</td>
<td>75.5</td>
<td>80.2</td>
<td>72.3</td>
<td>66.3</td>
<td>47.6</td>
<td>83.0</td>
<td>84.2</td>
<td>86.1</td>
<td>54.7</td>
<td>78.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>75.2</td>
<td>77.6</td>
<td>78.6</td>
<td>74.5</td>
<td>67.6</td>
<td>53.4</td>
<td>82.8</td>
<td>84.5</td>
<td>86.0</td>
<td>56.4</td>
<td>80.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the Faster R-CNN and SSD in Table 3 are from references [1, 3]. It can be seen from the detection results in the table that the proposed multi-scale feature fusion model obtains the best mAP. There has been a significant improvement in the detection performance on small objects such as birds, boats, chairs and plants. There have also been some improvements in the detection of medium and large objects such as airplanes, tables, and sheep. However, in contrast, the improvement on medium and large objects is obviously not as good as the improvement of small objects. There are two reasons for the improvement of detection performance. The first is using the oversampling strategy, increase the number of small objects samples, and get more information about small objects during training. The second reason is the fusion of multi-scale feature maps, which introduces the high semantic information low-resolution feature maps into high-resolution feature maps, which increases the accuracy of small objects detection.

![Figure 3](image-url). Comparison of detection results
In Figure 3, picture (a) and picture (d) are the original pictures to be detected. The cows in the distance in picture (a) and the persons in the distance in picture (b) occupy a small proportion in the whole picture and belong to small objects. Picture (b) and picture (e) are the results of the SSD. We can see that the cows in the distance in picture (b) and the persons in the distance in picture (e) are not detected, only cows, persons and bicycles with a large proportion were detected. Picture (c) and picture (f) are the results of our algorithm, which not only detects nearby cows, persons and bicycles, but also detects cows and persons with a small proportion in the distance. By comparison we can see that our algorithm is better than SSD in small objects detection.

5. Conclusions

In this paper, the SSD model is improved for the problem of weak performance of small objects detection. Exploiting the semantic information in low-resolution feature maps, multi-scale feature fusion is performed, and the sample numbers of small objects and large objects is balanced by the oversampling strategy. The object detection performance is better than Faster R-CNN and SSD. Since the deconvolution layer, the convolution layer, the Batch Normalization layer and the activation function layer are added in the process of multi-scale feature fusion, the number of parameters and the amount of calculation has increased, and the detection speed is lowered. In future research, we will focus on how to improve the efficiency of fusion, increase the speed of detection, and reduce the impact of introducing deconvolution.

References